規準化から相関係数へ

—— 三省堂教科書の方法を GRAPES で実現する (2) * ——

よしだ　はじめ† 

2012 年 2 月 29 日 改訂第 2 版　［第 1 版：2011 年 12 月 13 日＠千葉県数学教育協議会 高校サークル］

要旨

新課程数学 I の「データの分析」で扱われる相関係数の導入に際し，**規準化したデータの散布図から相関係数を定義する**という元・三省堂教科書の方法を **GRAPES** [1]を利用して，インタラクティブに示す.

1　相関係数の導入と規準化

2012 年度から高校数学 I で統計を必修内容とした指導要領が実施される.「データの分析」の相関の指導において，一般的な教科書では，共分散を求め，これを標準偏差の積で割ることで相関係数を導入している.

これに対して，今はなき三省堂教科書『数学 C』[2] では，相関係数を次のように導入していた.

(1) 対になっているデータから散布図[3]をかき，図から相関のようすを見る.（p.108）

(2) データを規準化し，その散布図と規準化したデータから相関係数を定義する.（p.109）

(3) 表計算ソフトでの計算方法と共に，相関係数 r は，

$$r = \frac{(xy \text{ の平均}) - (x \text{ の平均}) \cdot (y \text{ の平均})}{(x \text{ の標準偏差}) \cdot (y \text{ の標準偏差})}$$

と計算できることを示す.（p.131）

すなわち，次のような違いがある.

一般の教科書：元の量のまま共分散を求め，その値を標準偏差の積で割る.

三省堂教科書：元の量を規準化して[4]，その値から共分散を求める.

一般に相関関係を調べようとする 2 組のデータの変量は単位もスケールも異なる．また，散布図は関数グラフとは異なり x から y へという方向性は持たないので，２つの量のどちらを x 軸，どちらを y 軸にとってもかまわない．さらに，散布図はデータの点の位置で量を示すグラフであるから，グラフの基線は任意でよく，必要な範囲をとればよい．これらのことから，描かれる散布図によって相関の印象はかなり異なったものになる．図 1 ～ 3 の散布図は同じデータから作成したものである．[5]

グラフから図の印象によって相関を見るには変量を先に規準化したほうがよい．また，標準偏差で割ることの意味もわかりやすくなる．三省堂の教科書はこのような考えで展開されている.

規準化とは平均値を 0，標準偏差を 1 とした目盛りに変換することである．元のデータの x の平均値を m_x，標準偏差を s_x，また，y の平均値を m_y，標準偏差を s_y とすると，x に対しては $u = \dfrac{x - m_x}{s_x}$，　y に対しては $v = \dfrac{y - m_y}{s_y}$ という変換となる．x，y はそれぞれのデータの

---

*(1) は「指数関数の微分法と e の導入」（2008）

†吉田 一，河合塾 COSMO コース講師

[1]友田勝久氏によるグラフプレゼンテーションのフリーソフト．Web から入手可能.

[2]『数学 C』改訂版（2000），三省堂．1989 年指導要領のもとでの教科書．筆者は執筆者の一員.

[3]当時の教科書では「相関図」.

[4]この段階で標準偏差で割ることになる.

[5]データ：総務省統計局 Web なるほど統計学園 より

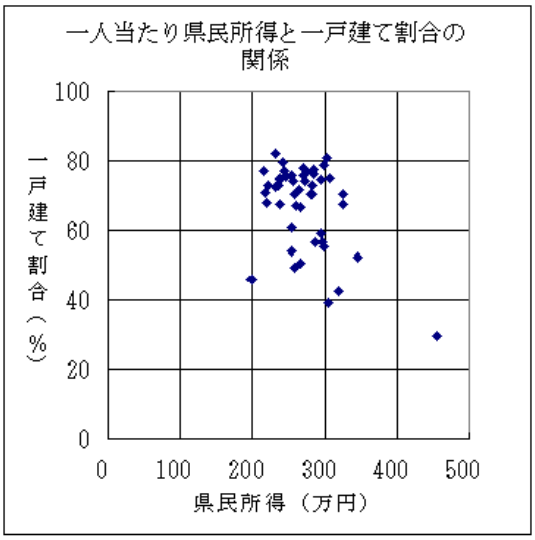

図 1　元の散布図

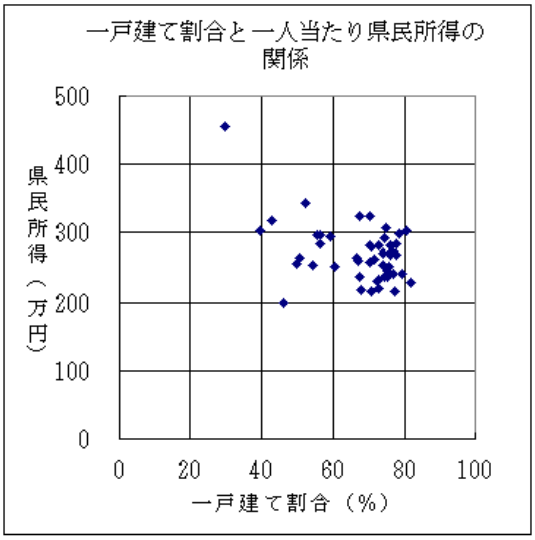

図 2　図 1 の軸を入れ替えた

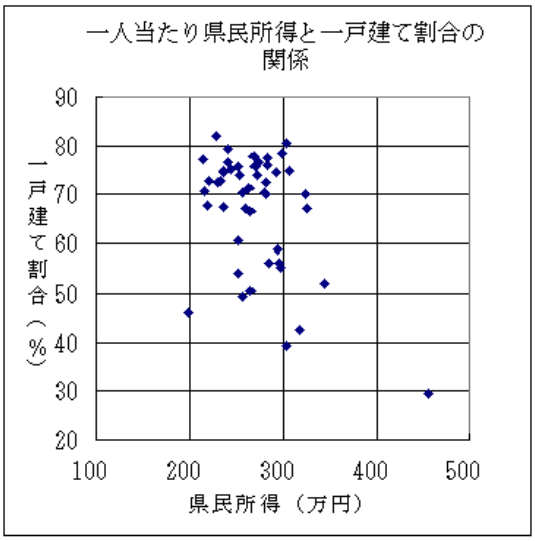

図 3　図 1 の目盛りの範囲を変えた

単位がついた量であるが，u，v は平均値から標準偏差の何倍離れているかを表す単位のない数になる．[6]

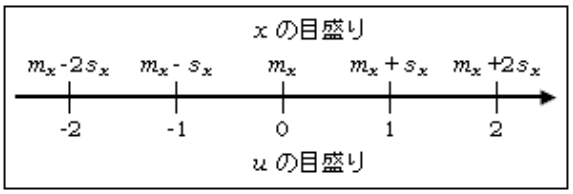

図 4　x 軸の元の目盛りと規準化した目盛り

これより，相関係数 r を

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - m_x}{s_x}\right)\left(\frac{y_i - m_y}{s_y}\right)$$

と定義する．[7]

規準化の計算は，例として少量のデータで行えばよい．規準化の意味を捉えさせるためにも，いきなり表計算ソフト等にたよらないほうがよいだろう．また，元のデータから直接共分散を求め，標準偏差の積で割る方法は後ほど扱うこととする．

さて，上記で述べた規準化は数値に対するものであるが，これを散布図に対して行うことを考える．つまり，元のデータに対する散布図を，x, y の平均値の組の点 $(m_x,\ m_y)$ を描画領域の中心とし，それぞれの標準偏差が等しい長さになるように目盛りをとって作成する．こうすることで，複数の散布図で相関の比較がしやすくなる．

具体的には，次のようにグラフ領域をとる．

- 散布図の描画領域を正方形にする．
- x の描画範囲を $m_x - ks_x \leqq x \leqq m_x + ks_x$，$y$ の描画範囲を $m_y - ks_y \leqq y \leqq m_y + ks_y$ とする．(k は正の定数)

k は正規分布であれば 4 程度，外れ値を考慮すると 5 程度に設定しておくとよいだろう．[8]

[6] 相関係数の指導の前に「偏差値」を取り上げれば，そこでこの変換が扱われることになる．

[7] 教科書（学習指導要領）では，数学 I では $\sum$ を使わないことになっているが，ここでは $\sum$ を使って記述した．

[8] 実際，図 1 ～ 3 のデータでは，$k = 4$ だとはみ出す点がでる．

2 GRAPES で規準化

散布図の規準化を **GRAPES** を用いて作成した.

目的 規準化の意味を示す. 規準化した散布図を表示する.

特徴 (1) 画面上のデータ（点）はドラッグで自由に動かすことができ，それに対応した相関係数の値を表示する. [9]

(2) x, y それぞれの平均値，標準偏差の範囲を画面上に示すことができる.

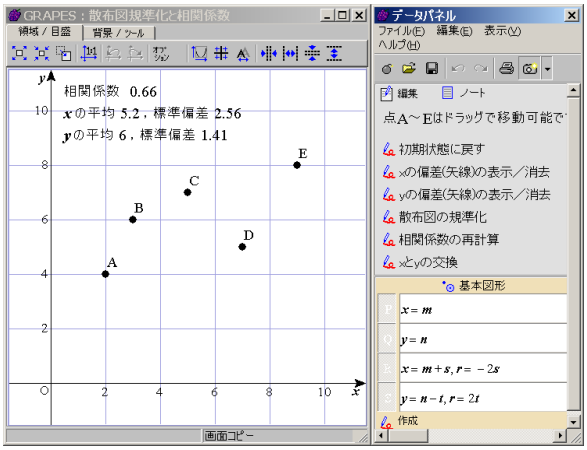

図 5 GRAPES の画面

散布図に平均値および標準偏差の範囲を表示した.（図 6）

さらに，平均値からの偏差を矢線で示した.（図 7）

散布図を規準化すると，図の中心に平均値がきて，両軸の ± 標準偏差の幅が等しく表示される.（図 8）

このように，散布図を規準化することによって，複数の散布図で図から相関の程度が比較しやすくなる. しかし，相関係数などの統計量は目安としての数値であり，図と一対一対応するものではないことは押さえておきたい.

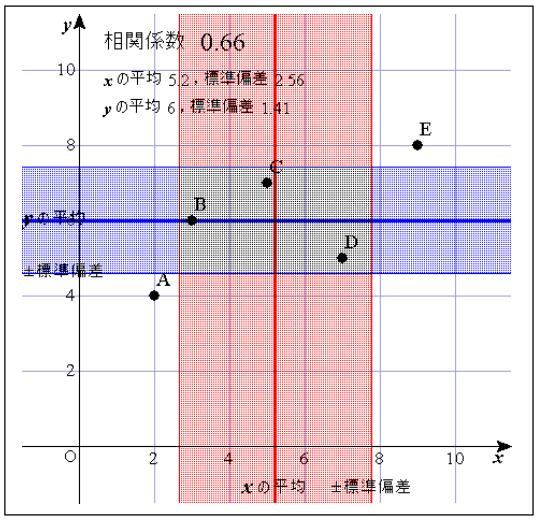

図 6 平均と標準偏差を表示

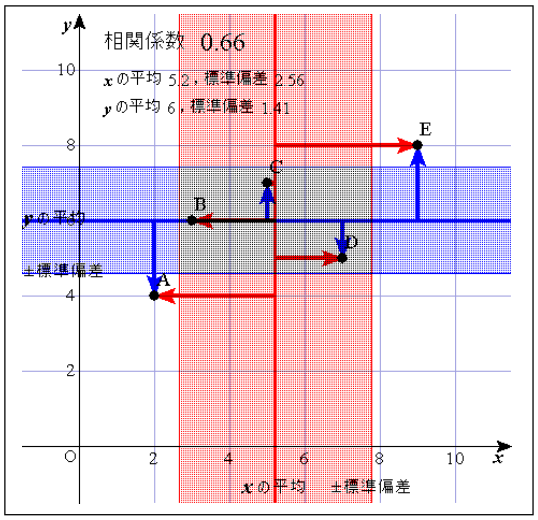

図 7 偏差の矢線を表示

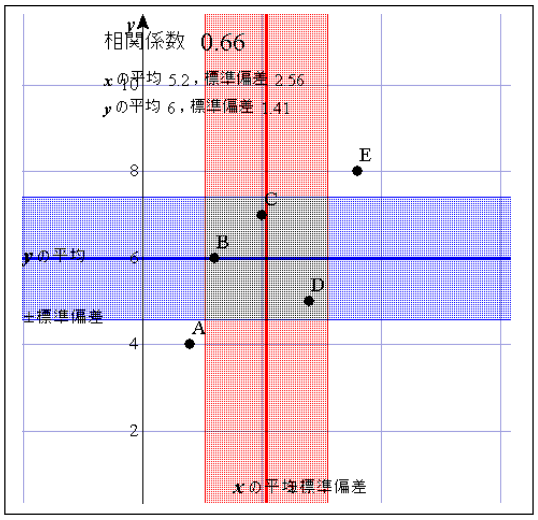

図 8 規準化した図

[9] 相関係数を教える前には，画面の相関係数の値は消しておくとよい.

図 1 のデータ（Excel の表）を **GRAPES** のテーブルにコピーし，散布図を描いた．(図 9) この場合はデータの点のドラッグはできない．x 軸，y 軸の目盛りは画面ではグラフの描画領域の下部，左部に描かれている．

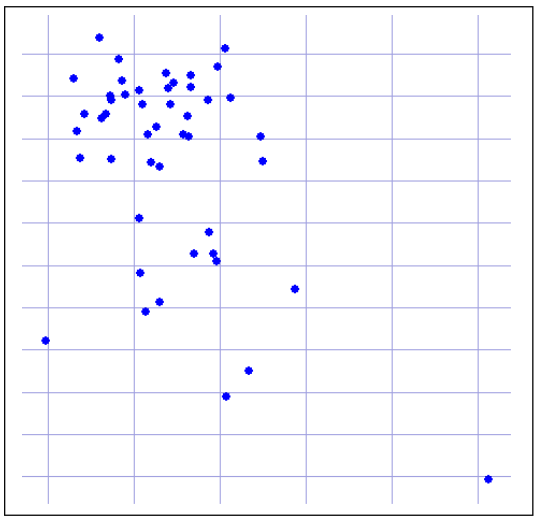

図 9　元データの散布図

このデータに対して規準化したグラフを表示させると，次のようになる．(図 10)

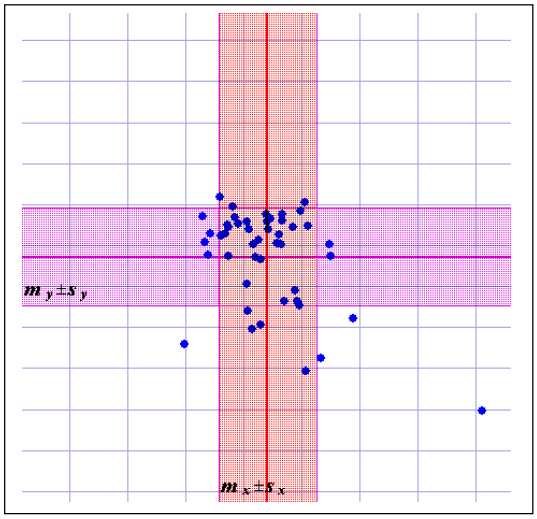

図 10　規準化した散布図

提供 GRAPES ファイル

1. 散布図規準化と相関係数.gps
   ドラッグ可能な散布図を規準化する．

2. 散布図規準化_実データ例.gps
   テーブルから散布図を描き，規準化する．

注意点（制限事項）

規準化すると，元の変量の値のとる範囲外の目盛りがふられることがある．たとえば，正の値しかとらないデータに負の目盛りまでとられるとか，割合に 100% を越える目盛りまでとられる，など．

【参考】相関係数の式変形

一般の教科書の定義への式変形は次のようになる．

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - m_x}{s_x}\right)\left(\frac{y_i - m_y}{s_y}\right)$$

$$= \frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y)}{s_x\, s_y}.$$

この分子は x の偏差と y の偏差の積の平均値であり，これはさらに次のように変形できる．

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y)\\
=&\frac{1}{n}\sum_{i=1}^{n}(x_i\, y_i - m_x\, y_i - m_y\, x_i + m_x\, m_y)\\
=&\frac{1}{n}\sum_{i=1}^{n} x_i\, y_i - m_x\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)\\
&\qquad - m_y\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + m_x\, m_y\\
=&\frac{1}{n}\sum_{i=1}^{n} x_i\, y_i - m_x\, m_y - m_y\, m_x + m_x\, m_y\\
=&\frac{1}{n}\sum_{i=1}^{n} x_i\, y_i - m_x\, m_y.
\end{aligned}$$

これは，**積の平均 − 平均の積** となっている．よって，相関係数は元のデータに対して

$$\frac{\text{積の平均} - \text{平均の積}}{\text{標準偏差の積}}$$

で求められ，一般の教科書の定義と同じになる．